Bayesian System Identification of a Nonlinear Dynamical System using a Novel Variant of Simulated Annealing

P.L.Green

Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield, United Kingdom, S1 3JD

Abstract

This work details the Bayesian identification of a nonlinear dynamical system using a novel MCMC algorithm: 'Data Annealing'. Data Annealing is similar to Simulated Annealing in that it allows the Markov chain to easily clear 'local traps' in the target distribution. To achieve this, training data is fed into the likelihood such that its influence over the posterior is introduced gradually - this allows the annealing procedure to be conducted with reduced computational expense. Additionally, Data Annealing uses a proposal distribution which allows it to conduct a local search accompanied by occasional long jumps, reducing the chance that it will become stuck in local traps. Here it is used to identify an experimental nonlinear system. The resulting Markov chains are used to approximate the covariance matrices of the parameters in a set of competing models before the issue of model selection is tackled using the Deviance Information Criterion.

Keywords: Bayesian model updating, Nonlinear system identification, Markov chain Monte Carlo, Simulated Annealing, Deviance Information Criterion.

Email address: p.l.green@sheffield.ac.uk (P.L.Green)

1. Introduction

This paper is concerned with the system identification of a nonlinear dynamical system using experimentally obtained training data. A probabilistic, Bayesian approach is utilised throughout. Such an approach is now well established in the structural dynamics community - relatively recent advances include the use of Bayesian methods in structural health monitoring [1], modal identification [2], state-estimation [3] (through use of the particle filter), the sensitivity analysis of large bifurcating nonlinear models [4] as well as an interesting study investigating the relations between frequentist and Bayesian approaches to probabilistic parameter estimation [5].

The identification problem detailed herein is one of *model selection* as well as *parameter estimation* such that, using experimental data $\mathcal{D}$, one must endeavor to find the optimum model $\mathcal{M}$ from a set of competing model structures as well as estimate the parameter vector $\boldsymbol{\theta}$ of that particular model. Using Bayes' theorem a measure of the plausibility of a parameter vector $\boldsymbol{\theta}$, given experimental data $\mathcal{D}$ and assumed model structure $\mathcal{M}$, is given by:

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})} \tag{1}$$

where $P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$ is the posterior probability density function (PDF) which one wishes to evaluate, $P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$ is termed the likelihood, $P(\boldsymbol{\theta}|\mathcal{M})$ the prior and $P(\mathcal{D}|\mathcal{M})$ the evidence. The likelihood represents the probability that the experimental training data $\mathcal{D}$ was witnessed according to the model $\mathcal{M}$ with parameters $\boldsymbol{\theta}$. Defining the likelihood requires the selection of an error-prediction model which describes the uncertainties present in the measurement and modelling processes (see [6] for a detailed discussion of error-prediction models). The prior is a PDF which represents one's parameter estimates for model $\mathcal{M}$ before the training data was known. The evidence is a normalising constant which ensures that the posterior PDF integrates to one.

This paper makes two main contributions. Firstly, a novel variant of Simulated Annealing (referred to as Data Annealing) is proposed and applied to a real system identification problem. It is shown to be computationally cheap and easy to tune. Secondly, it is shown that the issue of model selection of a real nonlinear dynamical system can be addressed using the Deviance Information Criterion (DIC). For the sake of readability the remainder of the introduction is split into two sections. The first outlines the motivation for the Data Annealing algorithm while the second focuses on the issue of model selection.

1.1. Motivation for the Data Annealing Algorithm

For the case where one is attempting to identify N_D parameters (such that $\boldsymbol{\theta} \in \mathbb{R}^{N_D}$), the evidence is given by:

$$P(\mathcal{D}|\mathcal{M}) = \int ... \int P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) P(\boldsymbol{\theta}|\mathcal{M}) d\theta_1 ... d\theta_{N_D}. \tag{2}$$

This integral is usually intractable and its multidimensional nature makes it too computationally expensive to evaluate numerically (if $N_D > 2$). Relatively early papers such as [7] made use of the property that the *maximum a posteriori* (MAP) parameter vector remains the same regardless of whether the posterior distribution has been normalised such that, through locating the MAP, a Taylor series expansion of the log posterior could be used to approximate the posterior PDF as a Gaussian[1]. Since then, an increase in computing power has allowed the adoption of Markov chain Monte Carlo (MCMC) methods. These involve the creation of an ergodic Markov chain whose stationary distribution is equal to the posterior PDF such that, once converged, the Markov chain is generating samples from $P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ (see [9] for more information on the convergence of Markov chains). This can be achieved without having to evaluate the evidence term. While many MCMC methods are available in the literature (Hamiltonian Monte Carlo for example [10]), by far the most popular is the Metropolis algorithm. Although well-established, a brief description of the Metropolis algorithm is given here as it helps to establish the motivation for the Data Annealing algorithm presented in Section 2 of this work.

Essentially, the aim of MCMC methods is to generate a sequence of samples $\left\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ...\right\}$ from a target PDF $\pi(\boldsymbol{\theta})/Z$ (where Z is a normalising constant). In the context of this paper, $\pi(\boldsymbol{\theta})$ represents the unnormalised posterior PDF and Z represents the evidence term. Initialising the Metropolis algorithm from parameter vector $\boldsymbol{\theta}^{(i)}$, a new state $\boldsymbol{\theta}'$ is proposed using a user-defined proposal PDF. The proposal PDF is conditional on the current state $\boldsymbol{\theta}^{(i)}$. For example, in the case where a Gaussian proposal is used then the new state is generated according to

$$\boldsymbol{\theta}' \sim \mathcal{N}(\boldsymbol{\theta}^{(i)}, \Sigma) \tag{3}$$

[1]For more information the reader may wish to consult the description of the Laplace approximation given in reference [8]

(where Σ is a user-defined covariance matrix). The new state is then accepted with probability:

$$a = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(i)})}\right\}. \tag{4}$$

If accepted then $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}'$ else $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$. This has the property that if the proposed state $\boldsymbol{\theta}'$ is in a region of higher probability density than the current state then it is always accepted. However, the Markov chain is also able to move into regions of lower probability density. One of the benefits of using such an acceptance rule is that the acceptance probability a can be computed without having to evaluate the evidence term. It can be shown that such an acceptance rule allows the chain to generate samples from $\pi(\boldsymbol{\theta})$ (for more information references [8] and [11] are recommended).

The advantages of using MCMC are numerous. Recalling that the purpose of system identification is usually to establish a reliable model which can be used to accurately and robustly predict the system's future response then, using the notation outlined in [12], one may want to predict a structural quantity of interest $h(\boldsymbol{\theta})$ using:

$$R = \int ... \int h(\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) d\theta_1 ... d\theta_{N_D}. \tag{5}$$

While evaluating equation (5) is difficult (for the same reason it is difficult to evaluate the evidence term), if one has used an MCMC algorithm to generate samples $\{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(M)}\}$ from the posterior parameter distribution then equation (5) can be approximated by:

$$R \approx \frac{1}{M} \sum_{i=1}^{M} h(\boldsymbol{\theta}^{(i)}). \tag{6}$$

Additionally, it has been shown that important information with regards to parameter correlations can be realised through the use of MCMC methods [13] (this is also demonstrated in Section 4 of the present work). However, MCMC also has its disadvantages. Before samples from the target distribution can be drawn in an effective manner, the Markov chain must converge on the globally optimum region of the parameter space. This region can be difficult to locate as it is often very concentrated relative to the size of one's prior distribution. Additionally, the Markov chain may become 'stuck' in a region of probability density which is not the global optimum. Throughout this paper these regions are referred to as 'local traps'.

The issue of local trapping led to the development of the Simulated Annealing algorithm

[14]. This involves the introduction of a factitious temperature[2] variable T such that, at high temperatures, the Markov chain is able to easily travel over local traps in the parameter space. The temperature variable is then reduced such that the fine details of the target distribution are gradually introduced - this is demonstrated graphically for a bimodal target PDF in Figure 1 (where π_T represents one's target distribution at temperature T). The rate at which T is reduced is commonly referred to as the *annealing schedule*.

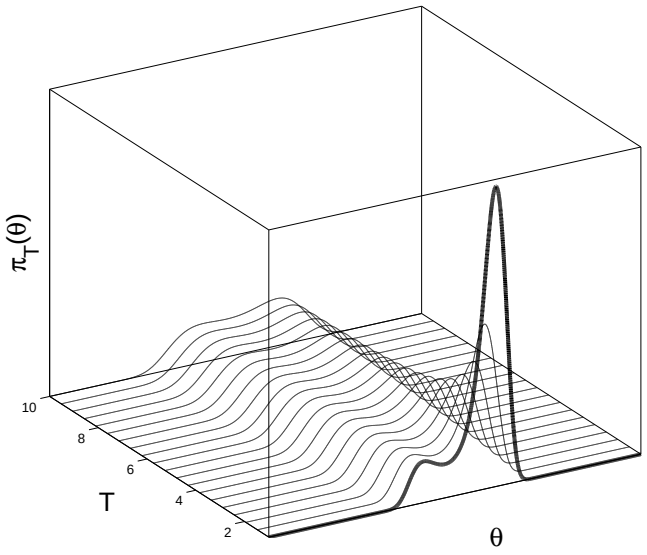

Figure 1: Graphical example of simulated annealing when $\boldsymbol{\theta} \in \mathbb{R}^1$.

Although this does not *guarantee* that the chain will converge on the optimum region of parameter space, Simulated Annealing has been established as a reliable optimisation algorithm. Soon after it was introduced several variants of Simulated Annealing were proposed [15, 16] in which the spread of the proposal PDF is initially set to be large but then reduces with temperature T (at a user-defined rate), thus encouraging the Markov chain to make large jumps at higher temperatures but conduct a more local search at lower temperatures.

When applied to Bayesian inference, the variable T can be introduced such that it controls the influence of the likelihood on the posterior:

[2]The phrases 'annealing' and 'temperature' are used as the Simulated Annealing algorithm was originally developed by drawing analogies with statistical physics [14]. The relations between Bayesian inference and statistical physics are discussed in [11].

$$\pi_T(\boldsymbol{\theta}) \propto P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})^T P(\boldsymbol{\theta}|\mathcal{M}). \tag{7}$$

Through using equation (7) as one's target distribution and defining an annealing schedule where T varies monotonically between 0 and 1, a gradual transition between the prior and posterior distribution can be realised. This concept was utilised in [12, 17, 18] where, by exploiting this gradual transition from prior to posterior, MCMC algorithms were developed which can be used to sample from posterior parameter distributions with complex geometries (where multiple, or even a continuum of optimum parameter vectors exist).

The performance of any Simulated Annealing algorithm will be sensitive to the choice of annealing schedule - annealing too fast places one at risk of becoming stuck in a local trap (such that a long time is required for the Markov chain to converge to its stationary distribution) while annealing too slowly will prove to be computationally expensive. It is possible to overcome this issue through the use of 'adaptive' annealing schedules such as those proposed in [17, 18, 19].

While the afore-mentioned algorithms are undoubtedly powerful, they can prove to be computationally expensive. One of the main aims of the current paper is to present a relatively cheap annealing algorithm which, within the context of Bayesian inference, can be applied to computationally demanding models.

1.2. Model Selection

The issue of model selection occurs when one must choose from a variety of competing model structures. This is complicated by the fact that models with more parameters will likely be able to better replicate some training data than models with less parameters. Consequently, if one judges models simply on their ability to replicate training data, then the most complex of the competing structures will always be accepted. Models which are overly-complex for the problem at hand are referred to as *overfitted*. Such models are often poor representations of the physics involved in the system of interest and, as a result, are poorly suited to making future predictions.

For a scenario where different model structures are available, the probability that the model $\mathcal{M}_i$ is suitable given the data $\mathcal{D}$ can also be written using Bayes' theorem:

$$P(\mathcal{M}_i|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathcal{D})} \tag{8}$$

thus allowing one to write the *relative* probability of two different models, given data $\mathcal{D}$, as:

$$\frac{P(\mathcal{M}_i|\mathcal{D})}{P(\mathcal{M}_j|\mathcal{D})} = \frac{P(\mathcal{D}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathcal{D}|\mathcal{M}_j)P(\mathcal{M}_j)} \tag{9}$$

where $P(\mathcal{M}_i)$ and $P(\mathcal{M}_j)$ represent one's prior beliefs in the suitability of each model (typically set equal to one another) and $P(\mathcal{D}|\mathcal{M})$ is the evidence term in equation (1). It is possible to show that the Bayesian approach to model selection automatically prevents overfitting (see [8] and [20] for more information). However, as has was described in the previous section, the evidence term is difficult to evaluate. As a result, one may instead choose to use a different model selection paradigm which is easier to evaluate than equation (9) but also retains the same model selection properties. In this work the Deviance Information Criterion (DIC) [21] is used as a model selection criterion.

Before describing the Deviance Information Criterion (DIC) it is convenient to first define the deviance:

$$D(\boldsymbol{\theta}) = -2\ln P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \tag{10}$$

where, as stated previously, $P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$ is the likelihood. The expected Deviance $E[D(\boldsymbol{\theta})]$ is a measure of how well the model structure $\mathcal{M}$ fits the data (as the parameter vector has been marginalised). The DIC is then defined as:

$$\text{DIC} = 2E[D(\boldsymbol{\theta})] - D(\hat{\boldsymbol{\theta}}). \tag{11}$$

where

$$E[D(\boldsymbol{\theta})] = \int P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})D(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{12}$$

and

$$\hat{\boldsymbol{\theta}} = E[P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})] = \int P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})\boldsymbol{\theta}d\boldsymbol{\theta} \tag{13}$$

such that the 'best' estimate parameters ($\hat{\boldsymbol{\theta}}$) are defined as the expected value of the posterior parameter distribution. Essentially, the lower the DIC the more favorable the model.

It also has the desired property that it rewards model fidelity while penalising model complexity (see reference [22] for a more detailed discussion).

The DIC lends itself well to situations where one has sampled from the posterior parameter distribution using MCMC as, using the successive parameter vectors realised by the MCMC algorithm $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(M)}\}$, the optimum parameter vector $\hat{\boldsymbol{\theta}}$ can be approximated by:

$$\hat{\boldsymbol{\theta}} \approx \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{\theta}^{(i)} \tag{14}$$

while the expected deviance can be also be approximated by:

$$E[D(\boldsymbol{\theta})] \approx \frac{1}{M} \sum_{i=1}^{M} D(\boldsymbol{\theta}^{(i)}) \tag{15}$$

thus allowing one to approximate the DIC. While this has been applied to synthetic data in [13], the current work demonstrates its application to real experimentally-obtained data.

The paper is organised as follows. In Section 2 the novel annealing algorithm is presented. In Section 3 the experimental system of interest is described. In Section 4 the results of the new annealing algorithm are analysed. This includes an analysis of the parameter correlations and predictive capabilities of competing model structures. The issue of model selection is then addressed using the Deviance Information Criterion (DIC). Section 5 is concerned with presenting possible future work while the conclusions are presented in Section 6.

2. Data Annealing

As stated in the previous section, MCMC methods can be used to generate samples from an unnormalised target PDF $\pi(\boldsymbol{\theta})$. In the context of this paper the target PDF is given by:

$$\pi(\boldsymbol{\theta}) = P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) P(\boldsymbol{\theta}|\mathcal{M}). \tag{16}$$

In practice it is usually desirable to evaluate the logarithm of the target PDF:

$$\ln(\pi(\boldsymbol{\theta})) = \ln(P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})) + \ln(P(\boldsymbol{\theta}|\mathcal{M})) \tag{17}$$

as, by first finding $\ln a = \ln \pi(\boldsymbol{\theta'}) - \ln \pi(\boldsymbol{\theta}^{(i)})$ before evaluating $a = \exp(\ln a)$), one can often avoid numerical overflow / underflow issues when calculating equation (4).

For the case where there are N measurements in the training data:

$$\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_N\} \tag{18}$$

then, assuming that each measurement is mutually independent, the likelihood is given by:

$$P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) = \prod_{i=1}^{N} P(\mathcal{D}_i|\boldsymbol{\theta}, \mathcal{M}). \tag{19}$$

In the case investigated here the training data $\mathcal{D}$ consists of a vector of inputs $\{y_1, y_2, ..., y_N\}$ and a vector of measured outputs $\{x_1, x_2, ..., x_N\}$ (the physical meaning of x and y are discussed in Section 3). Using a Gaussian error-prediction model allows the likelihood to be written as

$$P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) = \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \hat{x}_i(\boldsymbol{\theta}))^2 \right) \right] \tag{20}$$

where $\hat{x}_i(\boldsymbol{\theta})$ represents the response of the model with parameters $\boldsymbol{\theta}$ and σ^2 is the likelihood variance (which can be treated as another parameter to be found). Consequently, a single evaluation of the likelihood requires the simulation of N data points. It is suggested here that, rather than using T to control the influence of the likelihood on the posterior (as with Simulated Annealing), a similar effect can be achieved by varying the amount of data used in the likelihood. In other words, it is possible to increase the influence of the likelihood through the introduction of additional data points into $\mathcal{D}$. The rate at which the data points are introduced can be controlled according to a user-defined schedule - this is conceptually similar to the annealing schedule used in Simulated Annealing. The major advantage of this method is that it is computationally fast - in the early stages of the algorithm relatively few points need to be simulated by the model per evaluation of the likelihood. Throughout the current work this method is referred to as Data Annealing. It should be noted that the concept of annealing through the gradual addition of data points in the likelihood was proposed but not actually implemented in [12].

As was stated in Section 1, the Metropolis algorithm requires a user-defined proposal PDF to generate candidate parameter vectors $\boldsymbol{\theta}'$ - this is often chosen to be a Gaussian. In the current work the proposal PDF will be denoted $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)})$. In [15] it was suggested that, to reduce the probability of the Markov chain becoming stuck in a local trap, a proposal distribution with larger tails should used in place of a Gaussian distribution. Specifically, it was suggested that a Cauchy distribution could be utilised as, while it is locally similar to

a Gaussian, it possesses larger tails (as shown in Figure 2). This is desirable as, while the resulting Markov chain will spend the majority of the time conducting a local search of the parameter space, it will also occasionally propose relatively large jumps (thus increasing its ability to escape from local traps).

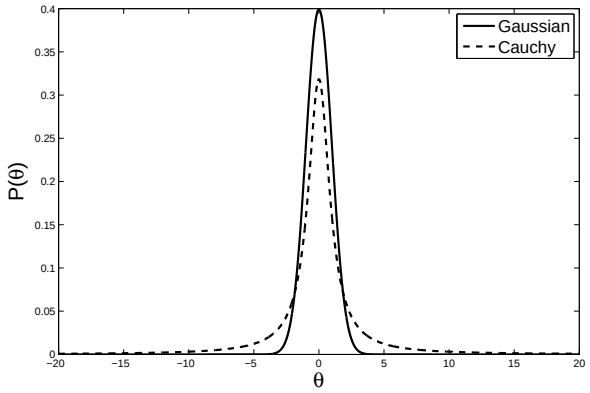

Figure 2: Comparison between Gaussian and Cauchy probability density functions.

A disadvantage of this method becomes apparent when the dimension of the parameter space is greater than one as samples from the multidimensional Cauchy distribution are not uncorrelated - large jumps in one parameter will often be accompanied by large jumps in all of the other parameters [11]. In the author's opinion this seems rather restrictive. Here, it is proposed that each parameter in $\boldsymbol{\theta}$ can be sampled independently from a one dimensional Cauchy distribution such that, for parameter θ_n:

$$q(\theta_n'|\theta_n^{(i)}) = \left[\pi\lambda_n\left(1+\left(\frac{\theta_n'-\theta_n^{(i)}}{\lambda_n}\right)^2\right)\right]^{-1} \tag{21}$$

(where λ_n controls the width of the distribution). Consequently, for the case where $\boldsymbol{\theta} \in \mathbb{R}^{N_D}$, the complete proposal distribution is simply the product of N_D Cauchy distributions:

$$q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)}) = \prod_{n=1}^{N_D} q(\theta_n'|\theta_n^{(i)}). \tag{22}$$

The result is a valid PDF which integrates to one, maintains the irreducibility of the Markov chain, allows one to perform a local search with occasional long jumps and does not have the afore-mentioned restrictive properties of the multidimensional Cauchy distribution. In fact, this property is so useful that an effective exploration of the parameter space can be achieved without having to vary the spread of the independent distributions $\{\lambda_1, ..., \lambda_{N_D}\}$ with annealing time - this is demonstrated in Section 4 of the current work. It should be

noted that in equation (22) one has the option of choosing different proposal widths for different parameters. This may be advantageous when the parameters are of very different scales. However, it was found here here that simply running the Data Annealing algorithm using the logarithm of the parameter vector allowed one to achieve good mixing despite using the same distribution width for each parameter.

3. Nonlinear System

A schematic of the nonlinear dynamical system of interest is shown in Figure 3. A 'centre magnet' is positioned such that it is free to slide along an aluminium rod via a set of linear bearings. Two 'outer magnets' are attached to the aluminium rod - they are positioned such that their poles oppose that of the centre magnet (thus creating a magnetic restoring force on the centre magnet). Consequently, when excited by the shaker, the centre magnet experiences oscillatory motion relative to the shaker table. Originally developed in the context of nonlinear energy harvesting, it is known that the magnetic restoring force on the centre magnet can be closely approximated using a linear and cubic stiffness term (similar to the hardening spring Duffing oscillator) [23]. As a result, the equation of motion of the system is:

$$m\ddot{x} = -c\dot{z} - kz - k_3 z^3 - mg - F, \qquad z = x - y \tag{23}$$

where x is the absolute displacement of the centre magnet, y is the displacement of the shaker table, m is the mass of the centre magnet, c is viscous damping, k is the linear stiffness, k_3 is the cubic stiffness and g is gravity. The training data $\mathcal{D}$ is made up of discretely sampled values of the excitation y (measured using the LVDT in Figure 3) and of the centre magnet response x (measured using the laser in Figure 3). The quantity F represents the force on the centre magnet as a result of friction effects. Three different friction models were considered. Firstly it was investigated whether the friction effects could be modelled simply using the viscous damping term c. Secondly, the Coulomb damping model was utilised such that:

$$F = F_c \operatorname{sgn}(\dot{z}) \tag{24}$$

where F_c is a parameter to be estimated. Finally, it was hypothesised that the hyperbolic tangent model was appropriate:

$$F = F_c \tanh(\beta\dot{z}) \tag{25}$$

(where F_c and β are parameters to be estimated). Throughout this paper these candidate models are referred to as the viscous, Coulomb and hyperbolic tangent models respectively. The hyperbolic tangent model has the property that

$$\lim_{\beta \to \infty} \tanh(\beta \dot{z}) = \text{sgn}(\dot{z}) \tag{26}$$

such that it is able to form a close approximation to the signum function without being discontinuous at $\dot{z} = 0$. It should be noted that the mass of the centre magnet was measured accurately before testing and so, in the following analysis, it is not included in the vector of parameters to be estimated.

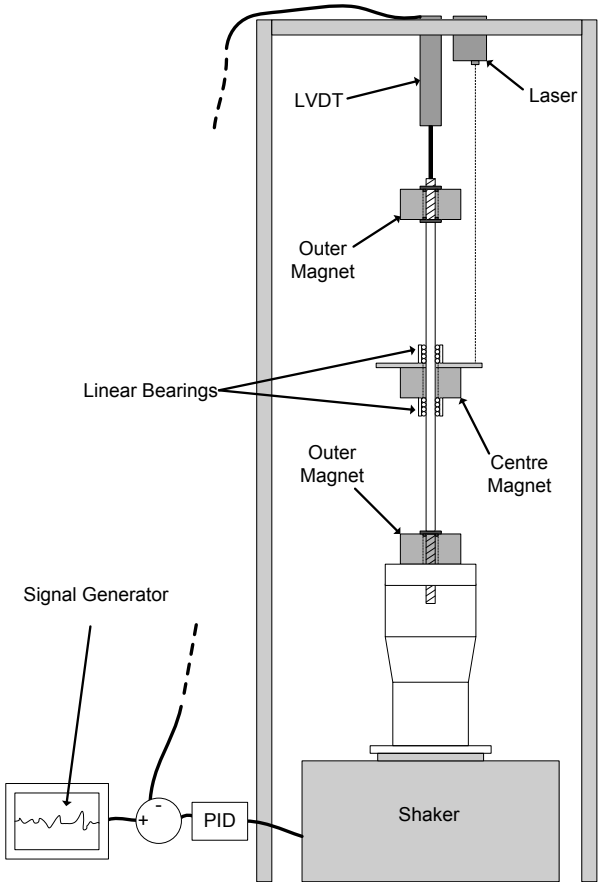

Figure 3: Schematic of experimental apparatus.

With regards to the applied excitation, a signal generator was used in conjunction with a PID controller to create a band-limited white noise acceleration. For a more detailed discussion of this experiment (which was also developed in the context of energy harvesting) the reader is directed towards references [24] and [25]. Two seconds of data measured at 1500 Hz was used as training data (this is shown in Figure 4).

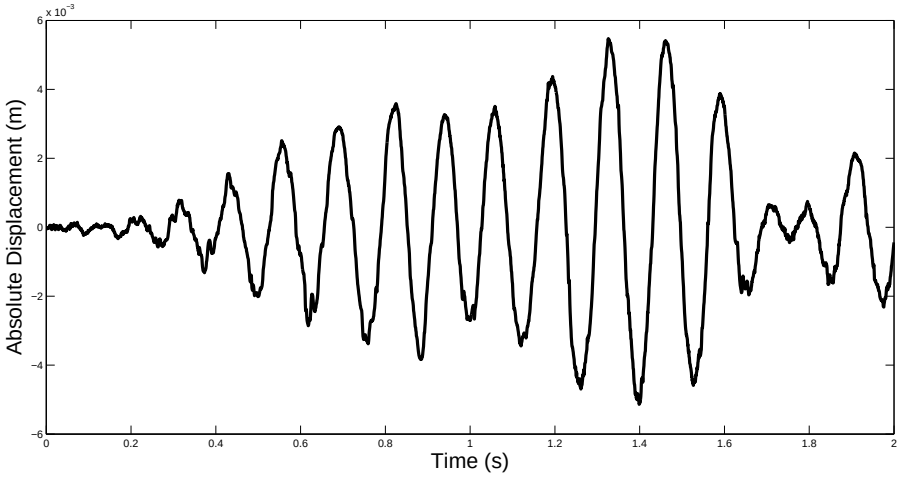

Figure 4: Two seconds of training data.

4. Results

4.1. Markov Chain Monte Carlo

Uniform (but not improper) prior distributions were used in all runs of the Data Annealing algorithm. The upper and lower limits of the priors for each parameter are shown in Table 1. A uncorrelated Gaussian error-prediction model (as described in Section 2) was used in the likelihood. It was assumed that the standard deviation of the likelihood (σ) was constant throughout the experimental test. In each of the following cases the value of σ was estimated alongside the other model parameters.

Parameter	Prior Lower Limit	Prior Upper Limit
c	0	0.2
F_c	0	0.01
β	0	1×10^7
k	0	80
k_3	0	1×10^7
σ	0	0.001

Table 1: Limits of uniform prior distribution.

For each model the Data Annealing algorithm was used to generate 50000 samples of $\boldsymbol{\theta}$. The proposal distribution shown in equation (22) was used with $\lambda = 0.005$ for each parameter. For the initial sample the data $\mathcal{D}$ used in the likelihood consisted of 2 points ($\{y_1, y_2\}$ and $\{x_1, x_2\}$). Additional data points were then introduced into the likelihood in a linear fashion for the first 2000 samples until the data $\mathcal{D}$ contained 3000 values of input (y) and 3000 values of the corresponding response (x). The amount of data $\mathcal{D}$ was then held constant for the remaining samples. The nonstationary portion of the resulting Markov chains were removed. To increase the independence between samples only every tenth sample from

the resulting Markov chain was used to approximate the marginal PDFs of the posterior distribution.

The resulting Markov chains and parameter histograms for the viscous damping, Coulomb and hyperbolic tangent models are shown in Figures 5, 6 and 7 respectively. As desired, use of the Data Annealing algorithm has allowed the Markov chain to make large jumps across the parameter space during the early stages while also allowing it to conduct a more local search once the chain has become stationary. To reiterate, this was achieved without having to vary the width of the proposal density.

With regards to Figure 7 it should be noted that the Markov chain for the β parameter did not appear to become stationary. This demonstrates an interesting flaw in the MCMC algorithm used in this paper: it is not clear whether the non-stationarity of the Markov chain is a result of β being a nuisance parameter or of a poorly tuned MCMC algorithm. Upon closer inspection it became apparent that at no point did the chain transition into a region lower than $\beta \approx 1000$. Recalling that the hyperbolic tangent model forms a close approximation to the Coulomb model when a large value of β is utilised allows one to hypothesise that the Coulomb model may be more appropriate in this case (the ability of all the models to predict future response and the issues of model selection are discussed in the subsequent sections).

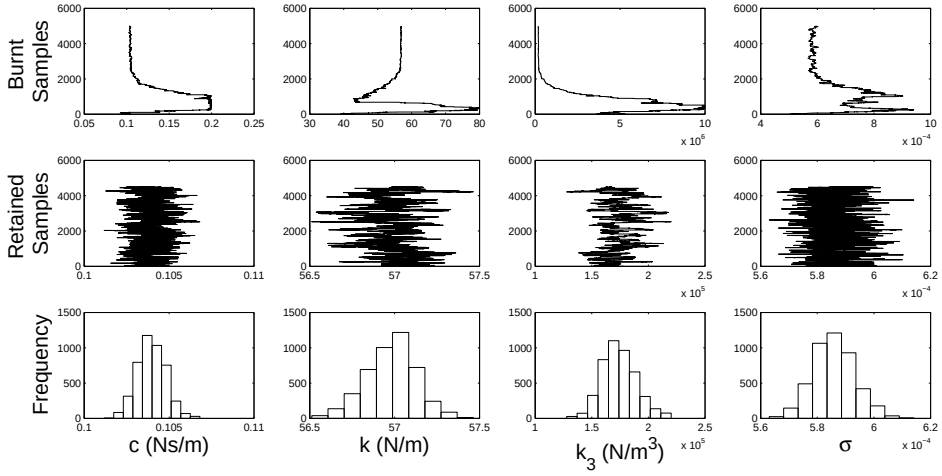

Figure 5: Results of the Data Annealing algorithm for the viscous model. The first row shows the burnt data during the annealing stage of the algorithm, the second row shows the thinned Markov chain with the burn period removed and the third row shows the resulting parameter histograms.

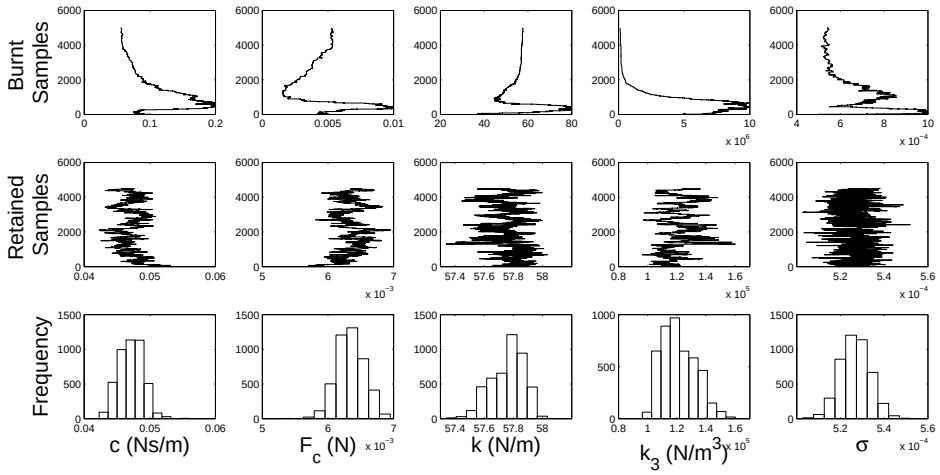

Figure 6: Results of the Data Annealing algorithm for the Coulomb model. The first row shows the burnt data during the annealing stage of the algorithm, the second row shows the thinned Markov chain with the burn period removed and the third row shows the resulting parameter histograms.

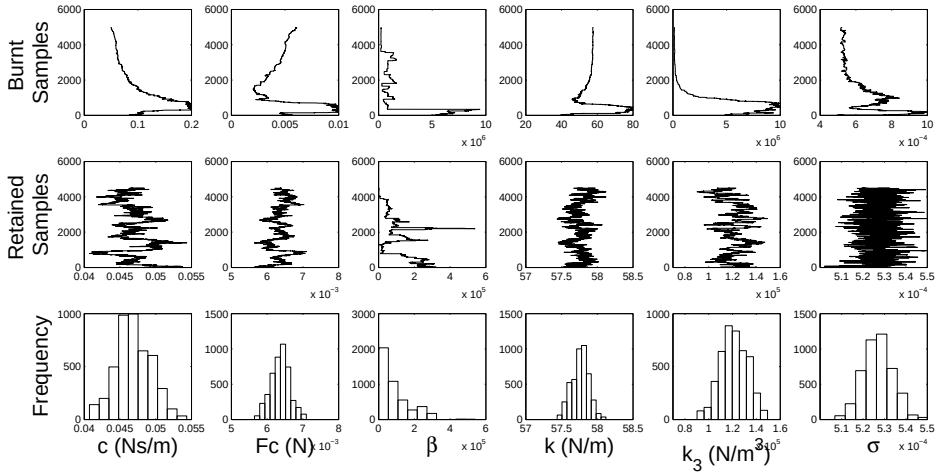

Figure 7: Results of the Data Annealing algorithm for the hyperbolic tangent model. The first row shows the burnt data during the annealing stage of the algorithm, the second row shows the thinned Markov chain with the burn period removed and the third row shows the resulting parameter histograms.

One of the advantages of using MCMC methods is that one can approximate the covariance matrix of the model parameters of a particular system. This is achieved by computing the correlation coefficients between the Markov chains of the different parameters. The resulting covariance matrices for the viscous, Coulomb and hyperbolic tangent models are shown in Figures 8, 9 and 10 respectively. For all three models it is interesting to note that there appears to be a strong negative correlation between the linear stiffness k and the nonlinear stiffness term k_3. This is a relation which is possible to show using the technique of equivalent linearisation: the situation where one is attempting to model the response of a system with a nonlinear hardening spring as accurately as possible using an equivalent

linear system. In such a case one must compensate for the lack of a nonlinear spring term via an increase in the linear spring term (see [26] for more details). In Figures 9 and 10 it is also shown that there is a strong negative correlation between the viscous damping term c, and F_c which controls the magnitude of friction in the system. This indicates that one may able to compensate for the lack of a friction model in a linear system through an increase in viscous damping. Again, this is something which can be shown using equivalent linearisation.

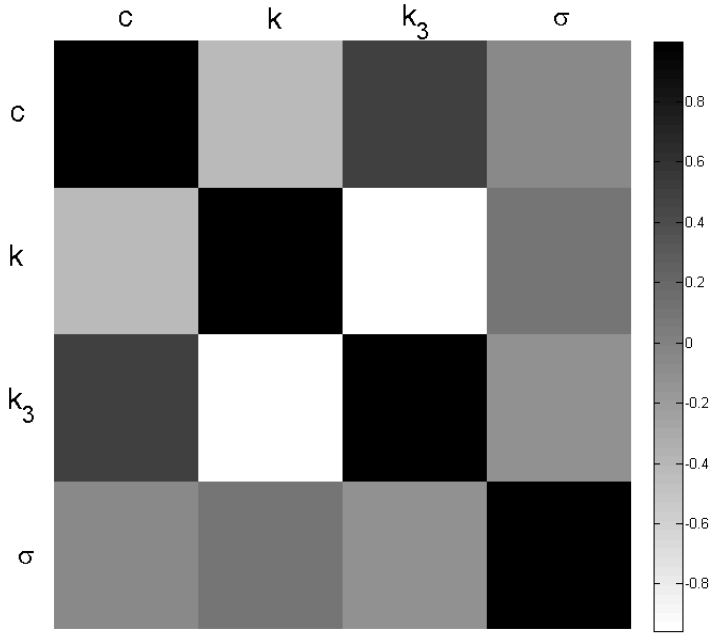

Figure 8: Covariance matrix for the viscous model.

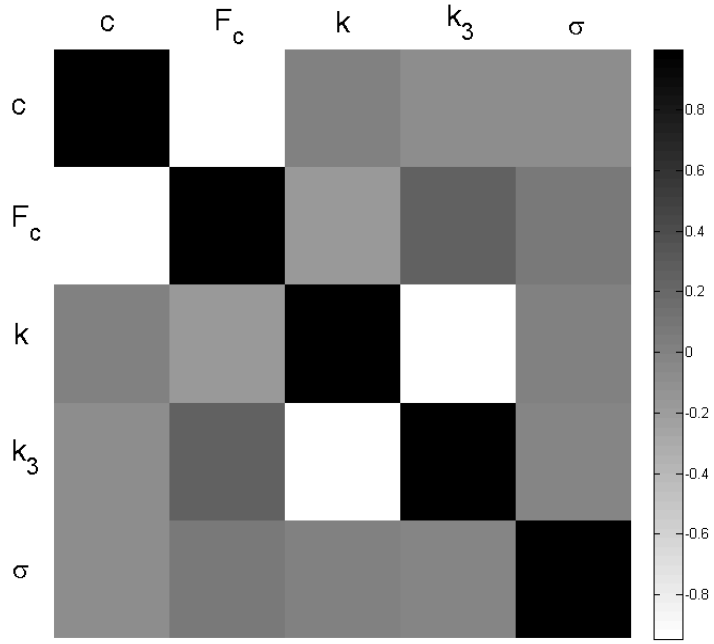

Figure 9: Covariance matrix for the Coulomb model.

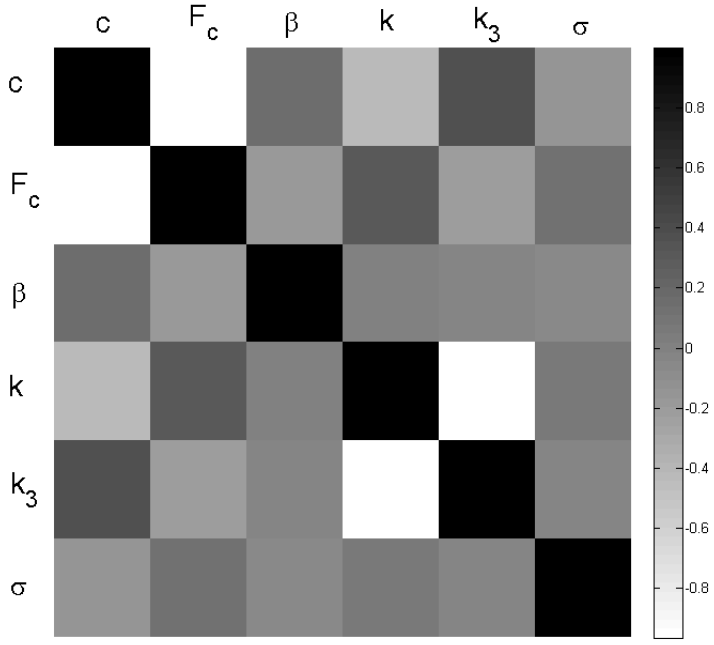

Figure 10: Covariance matrix for the hyperbolic tangent model.

4.2. Response Predictions

Having obtained probabilistic estimates for the parameters, each model was used to predict the response of the system to 59 seconds of a new excitation (which was part of a different set of experimental data). This data set will be denoted $\mathcal{D}_{new}$ to distinguish it

from the training data $\mathcal{D}$. As stated in [20], the Theorem of Total Probability can be used to obtain probabilistic estimates of $\mathcal{D}_{new}$:

$$P(\mathcal{D}_{new}|\mathcal{D},\mathcal{M}) = \int P(\mathcal{D}_{new}|\mathcal{D},\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})d\boldsymbol{\theta}. \tag{27}$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} P(\mathcal{D}_{new}|\boldsymbol{\theta}^{(i)},\mathcal{M}) \tag{28}$$

where $\boldsymbol{\theta}^{(i)}, i = 1, ..., M$ are the posterior samples generated by the Data Annealing algorithm.

An alternative method was suggested in [13] where, to account for the assumption that the system parameters are time-independent, it was suggested that one could sample a new parameter vector from the posterior *after every time step of the model simulation*. In the current work, both methods of uncertainty propagation were investigated (using an total ensemble of 50 model predictions) although it was found that the results were indistinguishable.

Figures 11 and 12 show the ability of the viscous and Coulomb models to replicate 1 second of the experimentally obtained response (with confidence bounds). It can be seen that both models have replicated the response of the system to a good level of accuracy. The prediction made by the hyperbolic tangent model is not shown here as it was indistinguishable from that of the Coulomb model. This strengthens the hypothesis that the Coulomb damping model is preferable to the hyperbolic tangent model as it is able to generate a very similar response despite having less parameters.

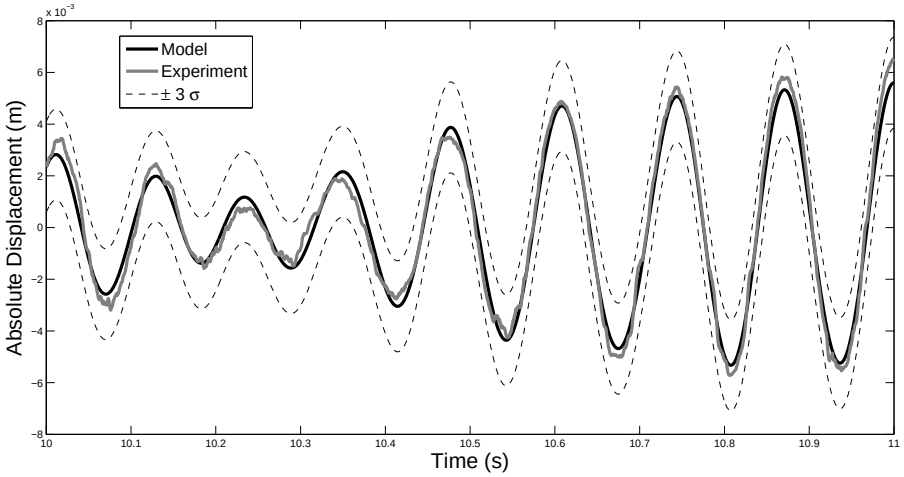

Figure 11: Comparison between one second of viscous model prediction (black) and one second of experimental data (grey) where dashed black lines represent 3 σ confidence bounds.

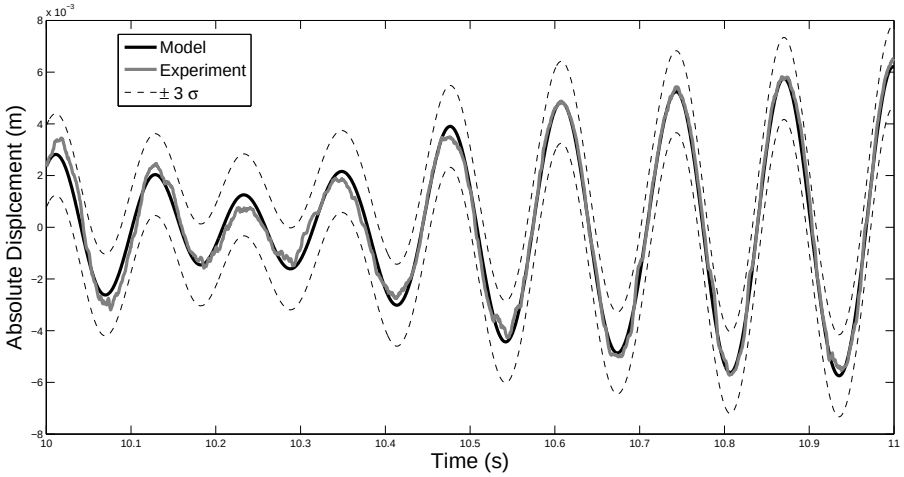

Figure 12: Comparison between one second of Coulomb model prediction (black) and one second of experimental data (grey) where dashed black lines represent 3 σ confidence bounds.

The mean square error (MSE) between the predicted future response from each model and the measured experimental response was calculated. This was taken over the entire 59 seconds of data. The MCMC samples realised in the previous section were then used to calculate the Deviance Information Criterion. The results are shown in Table 2. The MSE for the Coulomb and hyperbolic tangent models are significantly lower than that for the viscous model while the MSE for the Coulomb and hyperbolic tangent models are identical. This indicates that while the inclusion of a friction model has enhanced performance, the hyperbolic tangent model is simply acting as an approximation for the Coulomb model. This is confirmed by the Deviance Information Criterion which indicates that the Coulomb model is the most appropriate (thus confirming what was already suspected). For the sake of completeness, the ability of the Coulomb model to replicate the full 59 seconds of experimental data is shown in Figure 13.

Model	Parameter Number	MSE	DIC
Viscous	3	0.0175	-1.1047×10^6
Coulomb	4	0.0085	-1.1449×10^6
Hyperbolic Tangent	5	0.0085	1.3139×10^5

Table 2: Mean square error between model and experiment and Deviance Information Criterion for the viscous, Coulomb and hyperbolic tangent models.

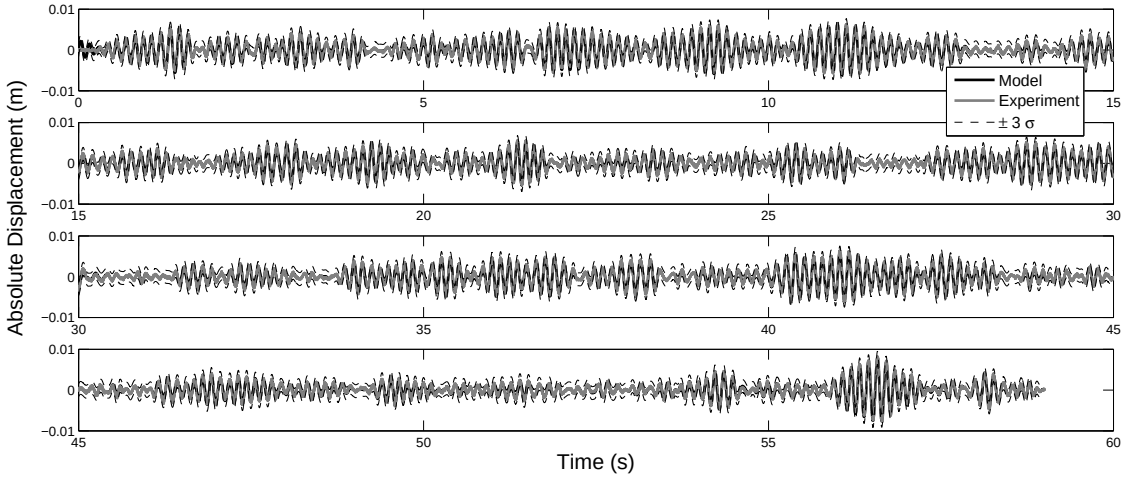

Figure 13: Comparison between 59 seconds of Coulomb model prediction (black) and fifty nine seconds of experimental data (grey) where dashed black lines represent 3 σ confidence bounds.

5. Discussion and Future Work

One of the disadvantages of Data Annealing is that, relative to algorithms such as Transitional MCMC (TMCMC) [17] and Asymptotically Independent Markov Sampling (AIMS) [18], the user has less control over the rate at which the influence of the likelihood is increased during the annealing process. This is because TMCMC and AIMS utilise the temperature variable in such a way that the transition from prior to posterior can be controlled in a continuous manner. The ability to select each temperature T from the set $T \in [0, 1]$ (subject to the constraint that the sequence of temperature values must increase monotonically from 0 to 1) essentially means that the user has an uncountably infinite set of possible annealing schedules available to them. This flexibility is lost when utilising the Data Annealing algorithm as the transition from prior to posterior is influenced by the sensitivity of one's parameter estimates to the introduction of a new data set. As a topic of future work the author aims to develop a version of Data Annealing algorithm which allows the user to have greater control over the annealing schedule.

Throughout this paper the DIC was used as a model selection criterion. The disadvantage of this approach is that, although it can be estimated using samples from the posterior, it is an *ad-hoc* penalty term which can only be used when each model has a single optimum parameter vector. A more complete approach would involve a variation of Data Annealing which was also able to estimate the model evidence (equation (2)) (thus allowing the relative plausibility of competing model structures to be investigated within a Bayesian framework). Consequently, for future work the author intends to investigate whether Data Annealing can

be combined with other MCMC methods which are capable of estimating the model evidence - such methods could include Simulated Tempering [27, 28], Reversible Jump MCMC [29], TMCMC [17], AIMS [18] and Nested Sampling [30].

6. Conclusions

In this paper the system identification of an experimental nonlinear dynamical system was investigated using three competing model structures. A new MCMC algorithm named 'Data Annealing' was proposed. Being conceptually similar to Simulated Annealing, Data Annealing is designed such that, at its initial stages, the prior distribution dominates the shape of the target distribution. This allows the Markov chain to move freely around the parameter space. Additional training data is then progressively introduced into the likelihood such that the influence of the likelihood on the posterior is gradually increased. This computationally cheap method improves the ability of the Markov chain to converge on the globally optimum region of the parameter space without getting stuck in 'local traps'. Additionally, the Data Annealing algorithm utilises a proposal distribution which allows it to conduct a local search of the parameter space accompanied by occasional long jumps. It was shown that this proposal distribution is well suited to the problem at hand as it initially allows the Markov chain to explore large regions of the parameter space while is also capable of providing a more local search once the chain has converged. This was achieved without having to alter the width of the proposal distribution. Having demonstrated the Data Annealing algorithm on a real system identification problem, the resulting Markov chains were used to extract approximate covariance matrices for all of the models investigated, thus revealing information about parameter correlations induced by the data. Finally, a model selection criterion known as the Deviance Information Criterion was used to select the most appropriate model from the set of competing structures. It was shown that the DIC can be used to identify a model which can accurately replicate a set of training data without being overfitted (relative to the other elements in a set of user-defined model structures).

7. Acknowledgements

The author would like to thank James L. Beck from the California Institute of Technology for his talk at IMAC XXXI which inspired much of the work shown in this paper.

This work was conducted as part of an EPSRC fellowship and is also closely aligned to the EPSRC Programme Grant 'Engineering Nonlinearity' EP/K003836/1.

[1] M.W. Vanik, J.L. Beck, and S.-K. Au. Bayesian Probabilistic Approach to Structural Health Monitoring. *Journal of Engineering Mechanics*, 126(7):738–745, 2000.

[2] K.-V. Yuen and L.S. Katafygiotis. Bayesian Fast Fourier Transform Approach for Modal Updating using Ambient Data. *Advances in Structural Engineering*, 6(2):81–95, 2003.

[3] J. Ching, J.L. Beck, and K.A. Porter. Bayesian State and Parameter Estimation of Uncertain Dynamical Systems. *Probabilistic Engineering Mechanics*, 21(1):81–96, 2006.

[4] W. Becker, K. Worden, and J. Rowson. Bayesian Sensitivity Analysis of Bifurcating Nonlinear Models. *Mechanical Systems and Signal Processing*, 34(1):57–75, 2013.

[5] S.-K. Au. Connecting Bayesian and Frequentist Quantification of Parameter Uncertainty in System Identification. *Mechanical Systems and Signal Processing*, 29:328–342, 2012.

[6] E. Simoen, C. Papadimitriou, and G. Lombaert. On Prediction Error Correlation in Bayesian Model Updating. *Journal of Sound and Vibration*, 332(18):4136–4152, 2013.

[7] J.L. Beck and L.S. Katafygiotis. Updating Models and their Uncertainties. I: Bayesian Statistical Framework. *Journal of Engineering Mechanics*, 124(4):455–461, 1998.

[8] D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[9] J.L. Doob. *Stochastic Processes*. Wiley publications in statistics. Wiley, 1953.

[10] S. Cheung and J.L. Beck. Bayesian Model Updating using Hybrid Monte Carlo Simulation with Application to Structural Dynamic Models with many Uncertain Parameters. *Journal of Engineering Mechanics*, 135(4):243–255, 2009.

[11] R.M. Neal. Probabilistic Inference using Markov Chain Monte Carlo Methods. Technical report, University of Toronto, 1993.

[12] J.L. Beck and S.-K. Au. Bayesian Updating of Structural Models and Reliability using Markov Chain Monte Carlo Simulation. *Journal of Engineering Mechanics*, 128(4):380–391, 2002.

[13] K. Worden and J.J. Hensman. Parameter Estimation and Model Selection for a Class of Hysteretic Systems using Bayesian Inference. *Mechanical Systems and Signal Processing*, 32:153–169, 2012.

[14] S. Kirkpatrick and M.P. Vecchi. Optimization by Simmulated Annealing. *Science*, 220(4598):671–680, 1983.

[15] H. Szu and R. Hartley. Fast Simulated Annealing. *Physics Letters A*, 122(34):157 – 162, 1987.

[16] L. Ingber. Very Fast Simulated Re-annealing. *Mathematical and Computer Modelling*, 12(8):967 – 973, 1989.

[17] J. Ching and Y.C. Chen. Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging. *Journal of Engineering Mechanics*, 133(7):816–832, 2007.

[18] J.L. Beck and K.M. Zuev. Asymptotically Independent Markov Sampling: a New Markov Chain Monte Carlo Scheme for Bayesian Inference. *International Journal for Uncertainty Quantification*, 3(5), 2013.

[19] P. Salamon, J.D. Nulton, J.R. Harland, J. Pedersen, G. Ruppeiner, and L. Liao. Simulated Annealing with Constant Thermodynamic Speed. *Computer Physics Communications*, 49(3):423–428, 1988.

[20] M. Muto and J.L. Beck. Bayesian Updating and Model Class Selection for Hysteretic Structural Models using Stochastic Simulation. *Journal of Vibration and Control*, 14(1-2):7–34, 2008.

[21] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[22] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003.

[23] B.P. Mann and N.D. Sims. Energy Harvesting from the Nonlinear Oscillations of Magnetic Levitation. *Journal of Sound and Vibration*, 319(1):515–530, 2009.

[24] P.L. Green, K. Worden, K. Atallah, and N.D. Sims. The Effect of Duffing-Type Non-Linearities and Coulomb Damping on the Response of an Energy Harvester to Random Excitations. *Journal of Intelligent Material Systems and Structures*, 23(18):2039–2054, 2012.

[25] P.L. Green, K. Worden, K. Atallah, and N.D. Sims. The Benefits of Duffing-Type Nonlinearities and Electrical Optimisation of a Mono-Stable Energy Harvester Under White Gaussian Excitations. *Journal of Sound and Vibration*, 331(20):4504–4517, 2012.

[26] K. Worden and G.R. Tomlinson. *Nonlinearity in Structural Dynamics: Detection, Identification and Modelling*. Taylor & Francis, 2010.

[27] E. Marinari and G. Parisi. Simulated Tempering: a New Monte Carlo Scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

[28] C.J. Geyer and E.A. Thompson. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

[29] P.J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732, 1995.

[30] J. Skilling. Nested Sampling for General Bayesian Computation. *Bayesian Analysis*, 1(4):833–859, 2006.